

Efficient moment-based inference of admixture parameters and sources of gene flow

Mark Lipson,^{1#} Po-Ru Loh,^{1#} Alex Levin,¹
David Reich,^{2,3} Nick Patterson,² Bonnie Berger^{1,2,*}

¹Department of Mathematics and Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge, MA 02139

²Broad Institute, Cambridge, MA 02142

³Department of Genetics, Harvard Medical School, Boston, MA 02115

[#]These authors contributed equally to this work.

^{*}To whom correspondence should be addressed:

Department of Mathematics 2-373

Massachusetts Institute of Technology

77 Massachusetts Avenue, Cambridge, MA 02139

Tel: (617) 253-1827; Fax: (617) 258-5429

E-mail: bab@mit.edu

Abstract

The recent explosion in available genetic data has led to significant advances in understanding the demographic histories of and relationships among human populations. It is still a challenge, however, to infer reliable parameter values for complicated models involving many populations. Here we present *MixMapper*, an efficient, interactive method for constructing phylogenetic trees including admixture events using single nucleotide polymorphism (SNP) genotype data. *MixMapper* implements a novel two-phase approach to admixture inference using moment statistics, first building an unadmixed scaffold tree and then adding admixed populations by solving systems of equations that express allele frequency divergences in terms of mixture parameters. Importantly, all features of the tree, including topology, sources of gene flow, branch lengths, and mixture proportions, are optimized automatically from the data and include estimates of statistical uncertainty. *MixMapper* also uses a new method to express branch lengths in easily interpretable drift units. We apply *MixMapper* to recently published data for HGDP individuals genotyped on a SNP array designed especially for use in population genetics studies, obtaining confident results for 30 populations, 20 of them admixed. Notably, we confirm a signal of ancient admixture in European populations—including previously undetected admixture in Sardinians and Basques—involving a proportion of 20-40% ancient northern Eurasian ancestry.

Introduction

The most basic way to represent the evolutionary history of species or populations is through a phylogenetic tree, a model that in its strict sense assumes that there is no gene flow between populations after they have diverged (Cavalli-Sforza and Edwards, 1967). In many settings, however, groups that have split from one another can still exchange genetic material. This is certainly the case for human population history, during the course of which populations have often diverged only incompletely or diverged and subsequently mixed again (Reich et al., 2009; Patterson et al., 2012; Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011; Pugach et al., 2011; Green et al., 2010; Reich et al., 2010). To capture these more complicated relationships, previous studies have considered models allowing for continuous migration among populations (Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011) or have extended simple phylogenetic trees into *admixture trees*, in which populations on separate branches are allowed to re-merge and form an admixed offspring population (Reich et al., 2009; Patterson et al., 2012; Chikhi et al., 2001; Wang, 2003; Sousa et al., 2009). Both of these frameworks, of course, still represent substantial simplifications of true population histories, but they can help capture a range of new and interesting phenomena.

Several approaches have previously been used to build phylogenetic trees incorporating admixture events. First, likelihood methods (Chikhi et al., 2001; Wang, 2003; Sousa et al., 2009) use a full probabilistic evolutionary model, which allows a high level of precision with the disadvantage of greatly increased computational cost. Consequently, likelihood methods can in practice only accommodate a small number of populations (Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011; Sirén et al., 2011). Moreover, the tree topology must generally be specified in advance, meaning that only parameter values can be inferred and not the arrangement of populations in the tree. By contrast, moment-based methods (Reich et al., 2009; Patterson et al., 2012) use only means and variances of allele frequency divergences. Moments are simpler conceptually and especially computationally, and they allow for more flexibility in model conditions. Their disadvantages can include reduced statistical power and difficulties in design-

ing precise estimators with desirable statistical properties (e.g., unbiasedness) (Wang, 2003). Finally, a number of studies have considered “phylogenetic networks,” which generalize trees to include cycles and multiple edges between pairs of nodes and can be used to model population histories involving hybridization (Huson and Bryant, 2006; Yu et al., 2012). However, these methods also tend to be computationally expensive.

In this work, we introduce *MixMapper*, a new computational tool that fits admixture trees by solving systems of moment equations involving the pairwise distance statistic f_2 (Reich et al., 2009; Patterson et al., 2012), which is the average squared allele frequency difference between two populations. The theoretical expectation of f_2 can be calculated in terms of branch lengths and mixture fractions of an admixture tree and then compared to empirical data. *MixMapper* can be thought of as a generalization of the *qpgraph* package (Patterson et al., 2012), which takes as input genotype data, along with a proposed arrangement of admixed and unadmixed populations, and returns branch lengths and mixture fractions that produce the best fit to allele frequency moment statistics measured on the data. *MixMapper*, by contrast, performs the fitting in two stages, first constructing an unadmixed scaffold tree via neighbor-joining and then automatically optimizing the placement of admixed populations onto this initial tree. Thus, no topological relationships among populations need to be specified in advance.

Our method is similar in spirit to the independently developed *TreeMix* method (Pickrell and Pritchard, 2012). Like *MixMapper*, *TreeMix* builds admixture trees from second moments of allele frequency divergences, although it does so via a composite likelihood maximization approach made tractable with a multivariate normal approximation. Procedurally, *TreeMix* is structured in a “top-down” fashion, whereby a full set of populations is initially fit as an unadmixed tree, and gene flow edges are added sequentially to account for the greatest errors in the fit (Pickrell and Pritchard, 2012). This format makes *TreeMix* well-suited to handling very large trees: the entire fitting process is automated and can include arbitrarily many admixture events simultaneously. In contrast, *MixMapper* is designed as an interactive tool to maximize flexibility and precision with a “bottom-up” approach, beginning with a carefully screened unadmixed

scaffold tree to which admixed populations are added with best-fitting parameter solutions.

We use *MixMapper* to model the ancestral relationships among 52 populations from the CEPH-Human Genome Diversity Cell Line Panel (HGDP) (Rosenberg et al., 2002; Li et al., 2008) using recently published data from a new, specially ascertained SNP array designed for population genetics applications (Keinan et al., 2007; Patterson et al., 2012). Previous studies of these populations have built simple phylogenetic trees (Li et al., 2008; Sirén et al., 2011), identified a substantial number of admixed populations with likely ancestors (Patterson et al., 2012), and constructed a large-scale admixture tree (Pickrell and Pritchard, 2012). Here, we add an additional level of quantitative detail, obtaining best-fit admixture parameters and bootstrap error estimates for 30 HGDP populations, of which 20 are admixed. The results include, most notably, a significant admixture event in the history of all sampled European populations (Patterson et al., 2012), among them Sardinians and Basques.

Methods

Problem setup and model assumptions

The basic problem we consider is as follows: given an array of genotype (i.e., SNP) data sampled from a set of individuals grouped by population, what can we infer about the phylogeny and admixture history of these populations? We assume that all SNPs are neutral, biallelic, and autosomal, and that divergence times are short enough that there are no double mutations. Thus, allele frequency variation—the signal that we harness—is governed entirely by genetic drift and admixture. We model admixture as a one-time exchange of genetic material: two parent populations mix to form a single descendant population whose allele frequencies are a weighted average of the parents'. This model is of course a very rough rendition of true mixture events, but it is flexible enough to serve as a reasonable first-order approximation and lends itself to efficient analysis using f -statistics (Reich et al., 2009; Patterson et al., 2012).

Constructing an unadmixed scaffold tree

Our *MixMapper* admixture-tree-building procedure consists of two phases (Figure 1), the first of which selects a set of unadmixed populations to use as a scaffold tree. We begin by computing f_3 statistics (Reich et al., 2009; Patterson et al., 2012) for all triples of populations P_1, P_2, P_3 in the data set and removing those populations P_3 with any negative values $f_3(P_3; P_1, P_2)$, which indicate admixture. We then use pairwise f_2 statistics to build a neighbor-joining tree on every subset of the remaining populations. In the absence of admixture, f_2 distances are additive along a phylogenetic tree (Text S1; cf. Patterson et al. (2012)), meaning that neighbor-joining should recover a tree with leaf-to-leaf distances that are completely consistent with the pairwise f_2 data (Saitou and Nei, 1987). We therefore evaluate the quality of each putative unadmixed tree according to its maximum error between fitted and actual pairwise distances. Because of model violation in real data, trees built on smaller subsets are more additive, but they are also less informative; in particular, it is beneficial to include populations from as many continental groups as possible in order to maximize the utility of the scaffold for admixture fitting. *MixMapper* provides a ranking of trees by additivity as a guide from which the user chooses a suitable unadmixed tree. Finally, *MixMapper* adjusts the chosen tree by re-optimizing its branch lengths (maintaining the topology inferred from neighbor-joining) to minimize the sum of squared errors of all pairwise f_2 distances.

Two-way admixture fitting

The second phase of *MixMapper* begins by attempting to fit additional populations independently as simple two-way admixtures between branches of the unadmixed tree (Figure 1). Assuming for the moment that we know the branches from which the ancestral mixing populations split for a given admixed population, we can construct a system of equations of f_2 statistics that allows us to infer parameters of the mixture (Text S1). Specifically, the squared allele frequency divergence between the admixed population and each unadmixed population X' can be expressed as an algebraic combination of known branch lengths along with four unknown mix-

ture parameters: the locations of the split points on the two parental branches, the combined terminal branch length, and the mixture fraction (Figure 2A). To solve for the four unknowns, we need at least four unadmixed populations X' that produce a system of four independent constraints on the parameters. This condition is satisfied if and only if the data set contains two populations X'_1 and X'_2 that branch from different points along the lineage connecting the divergence points of the parent populations from the unadmixed tree (Text S1). If the unadmixed tree contains $n > 4$ populations, we obtain a system of n equations in the four unknowns that in theory is dependent. In practice, the equations are in fact slightly inconsistent because of noise in the f_2 statistics and error in the point-admixture model, so we perform least-squares optimization to solve for the unknowns; having more populations helps reduce the impact of noise.

Algorithmically, *MixMapper* thus performs two-way admixture fitting by iteratively testing each pair of branches of the unadmixed tree as possible sources for the ancestral mixing populations. For each choice of branches, *MixMapper* builds the implied system of equations and finds the least-squares solution (under the constraints that unknown branch lengths are nonnegative and the mixture fraction α is between 0 and 1), ultimately choosing the pair of branches and mixture parameters producing the smallest residual norm. Our procedure for optimizing each system of equations uses the observation that upon fixing α , the system becomes linear in the remaining three variables (Text S1). Thus, we can optimize the system by performing constrained linear least squares within a basic one-parameter optimization routine over $\alpha \in [0, 1]$. To implement this approach, we applied MATLAB's `lsqlin` and `fminbnd` functions with a few auxiliary tricks to improve computational efficiency (detailed in the code).

Three-way admixture fitting

MixMapper also fits three-way admixtures, i.e., those for which one parent population is itself admixed (Figure 2B). Explicitly, after an admixed population M_1 has been added to the tree, *MixMapper* can fit an additional user-specified admixed population M_2 as a mixture between

the M_1 terminal branch and another (unknown) branch of the unadmixed tree. The fitting algorithm proceeds in a manner analogous to the two-way mixture case: *MixMapper* iterates through each possible choice of the third branch, optimizing each implied system of equations expressing f_2 distances in terms of mixture parameters. With two mixed populations, there are now $2n + 1$ equations— $f_2(M_1, X')$ and $f_2(M_2, X')$ for all unadmixed populations X' , and also $f_2(M_1, M_2)$ —and eight unknowns: two mixture fractions, α_1 and α_2 , and six linear branch length parameters (Figure 2B). Fixing α_1 and α_2 results in a linear system as before, so we perform the optimization using MATLAB’s `lsqlin` within `fminsearch` applied to α_1 and α_2 in tandem. The same mathematical framework extends to optimizing the placement of populations with arbitrarily many ancestral waves of admixture, but for simplicity and to reduce the risk of overfitting, we chose to limit this version of *MixMapper* to three-way admixtures.

Expressing branch lengths in drift units

All of the tree-fitting computations described thus far are performed using pairwise distances in f_2 units, which are mathematically convenient to work with owing to their additivity along a lineage (in the absence of admixture). However, f_2 distances are not directly interpretable in the same way as genetic drift D , which is a simple function of time and population size:

$$D \approx 1 - \exp(-t/2N_e) \approx 2 \cdot F_{ST},$$

where t is the number of generations and N_e is the effective population size (Nei, 1987). To convert f_2 distances to drift units, we apply a new formula, dividing the f_2 -length of each branch by a heterozygosity value that we infer for the ancestral population at the top of the branch (Text S2). Qualitatively speaking, this conversion corrects for the relative stretching of f_2 branches at different portions of the tree as a function of heterozygosity (Patterson et al., 2012). In order to infer ancestral heterozygosity values accurately, it is important to use SNPs that are ascertained in an outgroup to the populations involved, which we address further below.

Before inferring heterozygosities at ancestral nodes of the unadmixed tree, we must first determine the location of the root (which is neither specified by neighbor-joining nor involved in the preceding analyses). *MixMapper* does so by iterating through branches of the unadmixed tree, temporarily rooting the tree along each branch, and then checking for consistency of the resulting heterozygosity estimates. Explicitly, for each internal node P , we split its present-day descendants (according to the rerooted tree) into two groups G_1 and G_2 according to which child branch of P they descend from. For each pair of descendants, one from G_1 and one from G_2 , we compute an inferred heterozygosity at P (Text S2). If the tree is rooted properly, these inferred heterozygosities are consistent, but if not, there exist nodes P for which the heterozygosity estimates conflict. *MixMapper* is thus able to infer the location of the root as well as the ancestral heterozygosity at each internal node, after which it applies the drift length conversion as a post-processing step on fitted f_2 branch lengths.

Bootstrapping

In order to measure the statistical significance of our parameter estimates, we compute bootstrap confidence intervals (Efron, 1979; Efron and Tibshirani, 1986) for the inferred branch lengths and mixture fractions. Our bootstrap procedure is designed to account for both the randomness of the drift process at each of a finite number of SNPs and the random choice of individuals to represent each population. First, we divide the genome into 50 evenly-sized blocks, with the premise that this scale should easily be larger than that of linkage disequilibrium among our SNPs. Then, for each of 500 replicates, we resample the data set by (a) selecting 50 of these SNP blocks at random with replacement; and (b) for each population group, selecting a random set of individuals with replacement, preserving the number of individuals in the group.

For each replicate, we recalculate all pairwise f_2 distances and present-day heterozygosity values using the resampled SNPs and individuals (adjusting the bias-correction terms to account for the repetition of individuals) and then construct the admixture tree of interest. Even though the mixture parameters we estimate (branch lengths and mixture fractions) depend in

complicated ways on many different random variables, we can directly apply the nonparametric bootstrap to obtain confidence intervals (Efron and Tibshirani, 1986). For simplicity, we use a percentile bootstrap; thus, our 95% confidence intervals indicate 2.5 and 97.5 percentiles of the distribution of each parameter among the replicate trees.

Computationally, we parallelize *MixMapper*’s mixture-fitting over the bootstrap replicates using MATLAB’s Parallel Computing Toolbox.

Evaluating fit quality

We use several criteria to evaluate the mixture fits produced by *MixMapper* and distinguish high-confidence results from possible artifacts of overfitting.

First, we can compare *MixMapper* results to information obtained from other methods, such as the 3-population test (Reich et al., 2009; Patterson et al., 2012). Negative f_3 values for a given target population indicate robustly that the population is admixed, and comparing f_3 statistics for different reference pairs can give useful clues about the ancestral mixing populations. Thus, while the 3-population test relies on similar data to *MixMapper*, its simpler form makes it useful for confirmation.

Second, the consistency of parameter values over bootstrap replicates gives an indication of the robustness of the admixture fit in question. All of our results have some amount of associated uncertainty, but we dismiss those with particularly large error bars. Most often, this phenomenon is manifested in the placement of ancestral admixing populations: for poorly fitting admixtures, these will often move between different branches of the tree from one replicate to the next, signaling unreliable results.

Third, we place less faith in results with highly skewed values of α . We expect that if we try to fit a non-admixed population as an admixture, *MixMapper* should return a closely related population as the first branch with mixture fraction $\alpha \approx 1$ (and an arbitrary second branch). Indeed, we often observe this pattern, although we also find that the second branch tends to have a mixture fraction of at least a few percent. We expect overfitting is likely in these cases.

Fourth, for any inferred admixture event, the two mixing populations must be contemporaneous. Since we cannot resolve the three pieces of terminal drift lengths leading to admixed populations (Figure 2A) and our branch lengths depend both on population size and absolute time, we cannot say for sure whether this property is satisfied for any given mixture fit. In some cases, however, it is clear that no realization of the variables could possibly be consistent: if we infer an admixture between a very recent branch and a very old one with a small value of the terminal drift c , then we can confidently say the mixture is unreasonable.

Data set

We analyzed a SNP data set from 934 HGDP individuals grouped in 53 populations (Rosenberg et al., 2002; Li et al., 2008). Unlike most previous studies of the HGDP samples, however, we worked with recently published data generated using the new Affymetrix Axiom Human Origins Array (Patterson et al., 2012), which was designed with a simple ascertainment scheme for accurate population genetic inference (Keinan et al., 2007). It is well known that ascertainment bias can cause errors in estimated divergences among populations (Clark et al., 2005; Albrechtsen et al., 2010), since choosing SNPs based on their properties in modern populations induces non-neutral spectra in related samples. While there do exist methods to correct for ascertainment bias (Nielsen et al., 2004), it is much more desirable to work with *a priori* bias-free data, especially given that typical SNP arrays are designed using opaque ascertainment schemes.

To avoid these pitfalls, we used Panel 4 of the new array, which consists of 163,313 SNPs that were ascertained as heterozygous in the genome of a San individual (Keinan et al., 2007). This panel is special because there is evidence that the San are approximately an outgroup to all other modern-day human populations (Li et al., 2008; Gronau et al., 2011). Thus, while the Panel 4 ascertainment scheme distorts the San allele frequency spectrum, it is nearly neutral with respect to all other populations. In other words, we can think of the ascertainment as effectively choosing a set of SNPs (biased toward San heterozygosity) at the common ancestor of the remaining 52 populations, after which drift occurs in a bias-free manner. We excluded

61,369 SNPs that are annotated as falling between the transcription start site and end site of a gene in the UCSC Genome Browser database (Fujita et al., 2011). Most of the excluded SNPs are not within actual exons, but, as expected, the frequency spectra at these “gene region” loci were slightly shifted toward fixed classes relative to other SNPs, indicative of the action of natural selection (Figure S1). Since we assume neutrality in all of our analyses, we chose to remove these SNPs.

Results

Applying *MixMapper* to the HGDP populations, we built an unadmixed scaffold tree with 10 populations, upon which 20 more populations fit robustly as admixtures (Figure 3). We describe these results in detail below.

Unadmixed phylogeny inferred for 10 HGDP populations

The first phase of our *MixMapper* analysis identified an (approximately) unadmixed scaffold tree containing 10 populations: Dai, Japanese, Karitiana, Lahu, Mandenka, Naxi, Papuan, Suruí, Yi, and Yoruba (Figure 3B). Despite the focus of the HGDP on isolated populations, most of the 53 HGDP groups exhibit signs of admixture detectable by the 3-population test, as has been noted previously (Patterson et al., 2012). Our initial filtering step of removing populations with negative values f_3 values (indicative of recent admixture) left only 20 that are potentially unadmixed. Furthermore, most subsets including even half of those 20 populations exhibited significant divergence from the f_2 -additivity that should hold in the case of pure drift (Text S1; cf. Patterson et al. (2012)).

As mentioned above, it is desirable to include a wide range of populations in the scaffold tree to provide both geographic coverage and extra equations in order to facilitate the fitting of admixed populations. Additionally, including at least four continental groups provides a fairer evaluation of additivity; with three or fewer, any quartet of populations must contain

at least two that are closely related. At the same time, including too many populations can compromise accuracy. The populations we selected form one of the most additive 10-population subsets representing at least four of the five major continental groups (Africa, Americas, Asia, Europe, Oceania) in the HGDP data set. To check that their placement in an unadmixed tree is reasonable, we confirmed that none of the 10 populations can be fit in a reasonable way as an admixture on a tree built with the other nine (data not shown).

Ancient admixture in the history of present-day European populations

A notable feature of our unadmixed scaffold tree is that it does not contain any European populations. We had ruled out including any HGDP Europeans other than Sardinian and Basque on the basis of at least one significantly negative f_3 value, as has been previously reported (Patterson et al., 2012). Moreover, we found that potential f_2 -distance trees containing Sardinian or Basque along with representatives of at least three other continents were noticeably less additive than four-continent trees of the same size without Europeans. For example, on a set of 16 potentially unadmixed populations, none of the 100 most additive 10-population trees include Europeans. This points to the presence of admixture in Sardinian and Basque as well as the other European populations.

Using *MixMapper*, we added each of the European populations to the unadmixed tree via admixtures (Figure 4; Table 1). For all eight groups in the HGDP data set, the best fit was as a mixture of a population related to the common ancestor of Karitiana and Suruí (in varying proportions of about 20-40%, with Sardinian and Basque among the lowest and Russian the highest) with a population related to the common ancestor of all unadmixed non-African populations on the tree. All eight European populations were fit independently, but notably, their ancestors were found to branch from the scaffold tree at very similar points, suggesting a similar broad-scale history. Their branch positions are also qualitatively consistent with previous work that used the 3-population test to deduce ancient admixture for Europeans other than Sardinian and Basque (Patterson et al., 2012). To confirm the signal in Sardinian and Basque, we

re-computed their mixture proportions using f_4 ratio estimation (Reich et al., 2009; Patterson et al., 2012), which like *MixMapper* uses allele frequency statistics but in a very simple and intuitive framework. We estimated approximately 20-25% “ancient northern Eurasian” ancestry (Table S1), which is in very good agreement with our findings from *MixMapper* (Table 1).

At first glance, this inferred admixture might appear improbable on geographical and chronological grounds, but importantly, the two ancestral branch positions do not represent the mixing populations themselves. Rather, there may be substantial drift from the best-fit branches to the true mixing populations, indicated as branch lengths a , b , and c in Figure 4A. Unfortunately, these three lengths appear only in a fixed linear combination in the system of f_2 equations (Text S1), and current methods can only give estimates of this linear combination rather than the individual values (Patterson et al., 2012). One plausible arrangement, however, is shown in Figure 4A for the case of Sardinian.

Two-way admixtures outside of Europe

We also found several other populations that fit robustly onto the unadmixed tree using simple two-way admixtures (Table 2). All of these can be identified as admixed using the 3-population or 4-population tests (Patterson et al., 2012), but with *MixMapper*, we were able to provide the full set of best-fit parameter values to place them onto an admixture tree.

First, we found that four populations from North-Central and Northeast Asia—Daur, Hezhen, Oroqen, and Yakut—are likely descended from admixtures between native North Asian populations and East Asian populations related to Japanese. The first three are estimated to have roughly 10-30% North Asian ancestry, while Yakut has 50-70%. Melanesians fit optimally as a mixture of a Papuan-related population with an East Asian population close to Dai, in a proportion of roughly 80% Papuan-related, similar to previous estimates (Reich et al., 2011; Xu et al., 2012). Finally, we found that Han Chinese have an optimal placement as an approximately equal mixture of two ancestral East Asian populations, one related to modern Dai (likely more southerly) and one related to modern Japanese (likely more northerly), corroborating a previous

finding of admixture in Han populations between northern and southern clusters in a large-scale analysis of East Asia (HUGO Pan-Asian SNP Consortium, 2009).

Recent admixtures involving western Eurasians

Finally, we inferred the branch positions of several populations that are well known to be recently admixed (cf. Patterson et al. (2012); Pickrell and Pritchard (2012)) but for which one ancestral mixing population was itself anciently admixed in a similar way to Europeans. To do so, we exploited the capability of *MixMapper* to fit three-way admixtures (Figure 2B), using the anciently admixed branch leading to Sardinian as one ancestral source branch. First, we found that Mozabite, Bedouin, Palestinian, and Druze, in decreasing order of African ancestry, are all optimally represented as a mixture between an admixed western Eurasian population (not necessarily European) related to Sardinian and an African population (Table 3). We also obtained good fits for Uygur and Hazara as mixtures between a western Eurasian population and a population related to the common ancestor of all East Asians on the tree (Table 3).

Comparison to results from *TreeMix*

MixMapper provides a nuanced view of human population relationships, refining approximate phylogenetic tree models by incorporating numerous ancestral admixture events (Figure 3). Our results are similar in spirit to an admixture tree for HGDP populations produced with the *TreeMix* software (Pickrell and Pritchard, 2012) (Figure S2), but there are also important differences. Both methods fit Palestinian, Bedouin, Druze, Mozabite, Uygur, and Hazara as admixtures, but *MixMapper* analysis suggests that these populations are better-modeled as three-way admixed. *TreeMix* fits Brahui, Makrani, Cambodian, and Maya—all of which the 3-population test identifies as admixed but we are unable to place reliably with *MixMapper*—while *MixMapper* confidently fits Daur, Hezhen, Oroqen, Yakut, Melanesian, and Han. Perhaps most notably, *TreeMix*, unlike *MixMapper*, does not infer a widespread ancient admixture for Europeans, although it does identify gene flow from an ancestor of Native Americans to Russians and

from Orcadian to an ancestor of Americans, which could reflect the same historical signal (Figure S2).

MixMapper and *TreeMix* also differ in their coverage of populations ultimately modeled. With *MixMapper*, we chose to create admixture trees involving only pre-selected approximately unadmixed populations, upon which admixed populations of interest are added on a case-by-case basis only if they fit reliably as two- or three-way admixtures. In contrast, *TreeMix* returns a single large-scale admixture tree containing all populations in the data set, which may include some that can be shown to be admixed but are not modeled as such. Overall, we view *MixMapper* as “semi-automated” compared to *TreeMix*, which is almost fully automated. Both approaches have benefits: ours allows more manual guidance and lends itself to interactive use, whereas *TreeMix* requires less user intervention, although some care must be taken in choosing the number of gene flow events to include (10 in Figure S2) to avoid creating spurious mixtures. Finally, while *TreeMix* is quite efficient, running its HGDP analysis in a matter of hours, *MixMapper* runs analyses of individual populations on a faster time scale that promotes rapid interactive investigation. After initial setup to compute allele frequency statistics and potential scaffold trees, *MixMapper* determines the best-fit admixture model for a chosen population in a few seconds. We have found this aspect of the software useful for exploring the reliability of mixture fits and their sensitivity to assumptions; for example, it is easy to test very quickly the effect of tweaking the composition of the scaffold tree and to experiment with fitting populations of interest as two- or three-way admixtures.

Estimation of ancestral heterozygosity

A significant advantage of using data free from ascertainment bias is that it enables us to compute accurate estimates of the heterozygosity (over a given set of SNPs) throughout an unadmixed tree, including at ancestral nodes. This in turn allows us to convert branch lengths from f_2 units to easily interpretable drift lengths, as discussed above and in Text S2. In Figure 5C, we show our estimates for the heterozygosity (averaged over all San-ascertained SNPs used) at

the common ancestor of each pair of present-day populations in the tree. Consensus values are given at the nodes of Figure 5A. The imputed heterozygosity should be the same for each pair of descendants having the same common ancestor, and indeed, with the new data set, the agreement is excellent (Figure 5C). By contrast, inferences of ancestral heterozygosity are much less accurate using HGDP data from the original Illumina SNP array (Li et al., 2008) because of ascertainment bias (Figure 5B); f_2 statistics are also affected but to a lesser degree (Figure S3), as previously demonstrated (Patterson et al., 2012). We used these heterozygosity estimates to express branch lengths of all of our trees in drift units (see Text S2).

Discussion

The *MixMapper* framework generalizes and automates several previous moment-based admixture inference methods, incorporating them as special cases and enabling comprehensive testing across many potential admixture scenarios. Mathematically, *MixMapper* sets up and solves a system of equations relating unknown mixture parameters to a complete list of measured pairwise f_2 statistics. Methods such as the three-population test for admixture and f_4 ratio estimation (Reich et al., 2009; Patterson et al., 2012) have similar theoretical underpinnings, as they make inferences on small subsets of populations using f_3 and f_4 statistics, which can be expressed as linear combinations of f_2 statistics. The main benefit of *MixMapper* is that by analyzing more populations simultaneously and automatically considering different tree topologies and sources of gene flow, it produces results that can be easier to interpret and can uncover unexpected admixture events.

For example, negative f_3 values—i.e., three-population tests indicating admixture—can be expressed in terms of relationships among f_2 distances between populations in an admixture tree. In general, three-population tests can be somewhat difficult to interpret because the surrogate parent populations may not in fact be closely related to the true participants in the admixture, e.g., in the “outgroup case” (Reich et al., 2009; Patterson et al., 2012). The relations among the f_2 statistics incorporate this situation naturally, however, and solving the full system

recovers the true branch points wherever they are. As another example, f_4 ratio estimation infers mixture proportions of a single admixture event from f_4 statistics involving the admixed population and four unadmixed populations situated in a particular topology (Reich et al., 2009; Patterson et al., 2012). Whenever data for five such populations are available, the system of all f_2 equations that *MixMapper* solves to obtain the mixture fraction becomes equivalent to the f_4 ratio computation. More importantly, because *MixMapper* infers all of the topological relationships within an admixture tree automatically by optimizing the solution of the distance equations over all branches, we do not need to specify in advance where the admixture took place—which is not always obvious *a priori*. By using more than five populations, *MixMapper* also benefits from more data points to constrain the fit.

Due in part to this generality of the *MixMapper* approach, we were able to obtain the notable result that all European populations in the HGDP are optimally represented as mixtures between a population related to the common ancestor of Americans and a population related to the common ancestor of all non-African populations in our unadmixed tree, confirming and extending an admixture signal first reported by Patterson et al. (2012). Our interpretation is that most if not all modern Europeans are descended from at least one large-scale ancient admixture event involving, in some combination, at least one population of Mesolithic European hunter-gatherers; Neolithic farmers, originally from the Near East; and/or other migrants from northern or Central Asia. Either the first or second of these could be related to the “ancient western Eurasian” branch in Figure 4, and either the first or third could be related to the “ancient northern Eurasian” branch. Present-day Europeans differ in the amount of drift they have experienced since the admixture and in the proportions of the ancestry components they have inherited, but their overall profiles are similar.

Our results for Europeans are consistent with several previously published lines of evidence (Pinhasi et al., 2012). First, it has long been hypothesized, based on analysis of a few genetic loci (especially on the Y chromosome), that Europeans are descended from ancient admixtures (Semino et al., 2000; Dupanloup et al., 2004; Soares et al., 2010). Ancient European

admixture has been studied with genome-wide statistics in Patterson et al. (2012) with largely similar conclusions. Our results also suggest an interpretation for a previously unexplained *frappe* analysis of worldwide human population structure (using $K = 4$ clusters) showing that almost all Europeans contain a small fraction of American-related ancestry (Li et al., 2008). Finally, sequencing of ancient DNA has revealed substantial differentiation in Neolithic Europe between farmers and hunter-gatherers (Bramanti et al., 2009), with the former more closely related to present-day Middle Easterners (Haak et al., 2010) and southern Europeans (Keller et al., 2012; Skoglund et al., 2012) and the latter more similar to northern Europeans (Skoglund et al., 2012), a pattern perhaps reflected in our observed northwest-southeast cline in the proportion of “ancient northern Eurasian” ancestry (Table 1). Further analysis of ancient DNA may help shed more light on the sources of ancestry of modern Europeans.

One important new insight of our European analysis is that we detect the same signal of admixture in Sardinian and Basque as in the rest of Europe. As discussed above, unlike other Europeans, Sardinian and Basque cannot be confirmed to be admixed using the 3-population test (as in Patterson et al. (2012)), likely due to a combination of less “ancient northern Eurasian” ancestry and more genetic drift since the admixture (Table 1). The first point is further complicated by the fact that we have no unadmixed “ancient western Eurasian” population available to use as a reference; indeed, Sardinians themselves are often taken to be such a reference. However, *MixMapper* uncovered strong evidence for admixture in Sardinian and Basque through additivity-checking in the first phase and automatic topology optimization in the second phase, discovering an arrangement of unadmixed populations enabling admixture parameter inference, which we then verified directly with f_4 ratio estimation. Perhaps the most convincing evidence of the robustness of this inference is the fact that *MixMapper* infers branch points for the ancestral mixing populations that are very similar to those of other Europeans (Table 1), a concordance that is most parsimoniously explained by a shared history of ancient admixture among Sardinian, Basque, and other European populations. Finally, we note that because we fit all European populations directly onto our scaffold tree without assuming Sardinian or Basque to be

an unadmixed reference, our estimates of the “ancient northern Eurasian” ancestry proportions in Europeans are larger than those in Patterson et al. (2012) and we believe more accurate than those previously reported (Skoglund et al., 2012; Patterson et al., 2012).

It is worth noting that of the 52 populations (excluding San) in the HGDP data set, there were 22 that we were unable to fit in a reasonable way either on the unadmixed tree or as admixtures. In part, this was because our simple point-admixture model is intrinsically limited in its ability to capture complicated population histories. Most parts of the world have surely witnessed low levels of inter-population migration over time, especially between nearby populations, making it difficult to fit admixture trees to the data. We also found cases where having data from more populations would help the fitting process, for example for three-way admixed populations such as Maya where we do not have a sampled group with a simpler admixture history that could be used to represent two of the three components. Similarly, we found that while Central Asian populations such as Burusho, Pathan, and Sindhi have clear signals of admixture from the 3-population test, they likely have ancestry from several different sources (including sub-Saharan Africa in some instances), making them difficult to fit with *MixMapper*. Finally, we have chosen here to disregard admixture with archaic humans, which is known to be a small but noticeable component for most populations in the HGDP (Green et al., 2010; Reich et al., 2010).

In certain applications, full genome sequences are beginning to replace more limited genotype data sets such as ours, but we believe that our methods and SNP-based inference more generally will still be valuable in the future. Despite the increasing feasibility of sequencing, it is still much easier and less expensive to genotype samples using a SNP array, and with over 100,000 loci, the data used in this study provide substantial statistical power. Additionally, sequencing technology is currently more error-prone, which can lead to biases in allele frequency-based statistics (Pool et al., 2010): for example, rare alleles can be difficult to distinguish from incorrect base calls, meaning that error correction will tend to flatten empirical frequency spectra. We expect *MixMapper* will continue to contribute to an important niche of population history inference methods based on SNP allele frequency data.

Software

Source code for the *MixMapper* software is available at <http://groups.csail.mit.edu/cb/mixmapper/>.

Acknowledgments

We would like to thank George Tucker and Joe Pickrell for helpful discussions and suggestions. This work was supported by the National Science Foundation (Graduate Research Fellowship support to M.L., P.L., and A.L. and HOMINID grant #1032255 to D.R. and N.P.) and the National Institutes of Health (grant GM100233 to D.R. and N.P.)

References

- Albrechtsen A, Nielsen F, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*. 27:2534–2547.
- Bramanti B, Thomas M, Haak W, et al. (11 co-authors). 2009. Genetic discontinuity between local hunter-gatherers and Central Europe’s first farmers. *Science*. 326:137–140.
- Cavalli-Sforza L, Edwards A. 1967. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*. 19:233–257.
- Chikhi L, Bruford M, Beaumont M. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*. 158:1347–1362.
- Clark A, Hubisz M, Bustamante C, Williamson S, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*. 15:1496–1502.
- Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. 2004. Estimating the impact of prehistoric admixture on the genome of Europeans. *Molecular Biology and Evolution*. 21:1361–1372.

- Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics*. 7:1–26.
- Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*. 1:54–75.
- Fujita P, Rhead B, Zweig A, et al. (11 co-authors). 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*. 39:D876–D882.
- Gravel S, Henn B, Gutenkunst R, et al. (11 co-authors). 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*. 108:11983–11988.
- Green R, Krause J, Briggs A, et al. (11 co-authors). 2010. A draft sequence of the Neandertal genome. *Science*. 328:710–722.
- Gronau I, Hubisz M, Gulko B, Danko C, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*. 43:1031–1034.
- Haak W, Balanovsky O, Sanchez J, et al. (11 co-authors). 2010. Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biology*. 8:e1000536.
- HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science*. 326:1541–1545.
- Huson D, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*. 23:254–267.
- Keinan A, Mullikin J, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*. 39:1251–1255.
- Keller A, Graefen A, Ball M, et al. (11 co-authors). 2012. New insights into the Tyrolean Iceman’s origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*. 3:698.

- Laval G, Patin E, Barreiro L, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*. 5:e10284.
- Li J, Absher D, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 319:1100–1104.
- Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press.
- Nielsen R, Hubisz M, Clark A. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*. 168:2373–2382.
- Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics*. 192:1065–1093.
- Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*. 8:e1002967.
- Pinhasi R, Thomas M, Hofreiter M, Currat M, Burger J. 2012. The genetic history of Europeans. *Trends in Genetics*. 28:496–505.
- Pool J, Hellmann I, Jensen J, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Research*. 20:291–300.
- Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biology*. 12:R19.
- Reich D, Green R, Kircher M, et al. (11 co-authors). 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 468:1053–1060.
- Reich D, Patterson N, Kircher M, et al. (11 co-authors). 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*. 89:516–528.

- Reich D, Thangaraj K, Patterson N, Price A, Singh L. 2009. Reconstructing Indian population history. *Nature*. 461:489–494.
- Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M. 2002. Genetic structure of human populations. *Science*. 298:2381–2385.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 4:406–425.
- Semino O, Passarino G, Oefner P, et al. (11 co-authors). 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science*. 290:1155–1159.
- Sirén J, Marttinen P, Corander J. 2011. Reconstructing population histories from single nucleotide polymorphism data. *Molecular Biology and Evolution*. 28:673–683.
- Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert M, Götherström A, Jakobsson M. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*. 336:466–469.
- Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H, Torroni A, Richards M. 2010. The archaeogenetics of Europe. *Current Biology*. 20:R174–R183.
- Sousa V, Fritz M, Beaumont M, Chikhi L. 2009. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics*. 181:1507–1519.
- Wall J, Lohmueller K, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution*. 26:1823–1827.
- Wang J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*. 164:747–765.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L, et al. (6 co-authors). 2012. Genetic dating indicates that the Asian–Papuan admixture through eastern Indonesia corresponds to the Austronesian expansion. *Proceedings of the National Academy of Sciences*. 109:4574–4579.

Yu Y, Degnan J, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*. 8:e1002660.

Figures

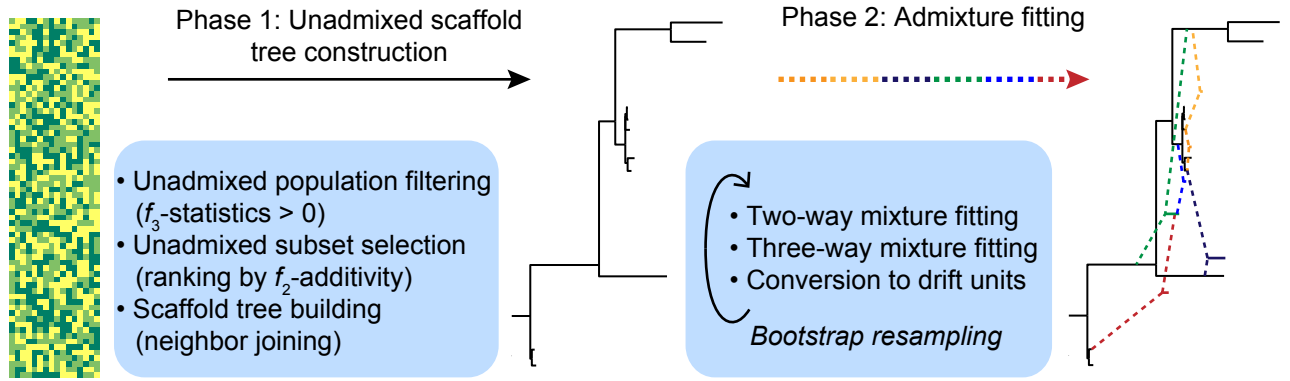


Figure 1. *MixMapper* workflow. *MixMapper* takes as input an array of SNP calls annotated with the population to which each individual belongs. The method then proceeds in two phases, first building a tree of (approximately) unadmixed populations and then attempting to fit the remaining populations as admixtures. In the first phase, *MixMapper* produces a ranking of possible unadmixed trees in order of deviation from f_2 -additivity; based on this list, the user selects a tree to use as a scaffold. In the second phase, *MixMapper* tries to fit remaining populations as two- or three-way mixtures between branches of the unadmixed tree. In each case *MixMapper* produces an ensemble of predictions via bootstrap resampling, enabling confidence estimation for inferred results.

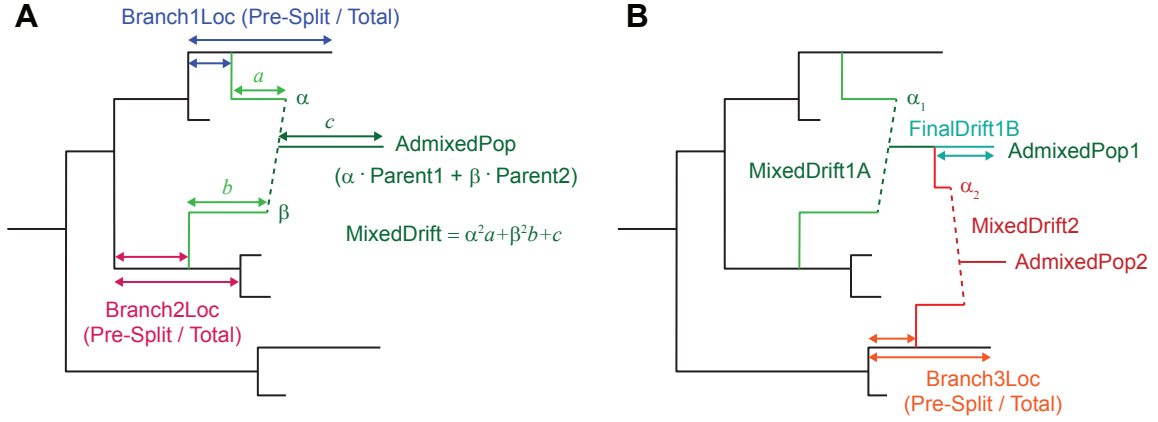


Figure 2. Schematic of mixture parameters fit by *MixMapper*. (A) A simple two-way admixture. *MixMapper* infers four parameters when fitting a given population as an admixture. It finds the optimal pair of branches between which to place the admixture and reports the following: Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches; α is the proportion of ancestry from Branch1 and $\beta = 1 - \alpha$ is the proportion from Branch2; and MixedDrift is the linear combination of drift lengths $\alpha^2 a + \beta^2 b + c$. (B) A three-way mixture: here AdmixedPop2 is modeled as an admixture between AdmixedPop1 and Branch3. There are now four additional parameters; three are analogous to the above, namely, Branch3Loc, α_2 , and MixedDrift2. The remaining degree of freedom is the position of the split along the AdmixedPop1 branch, which divides MixedDrift into MixedDrift1A and FinalDrift1B.

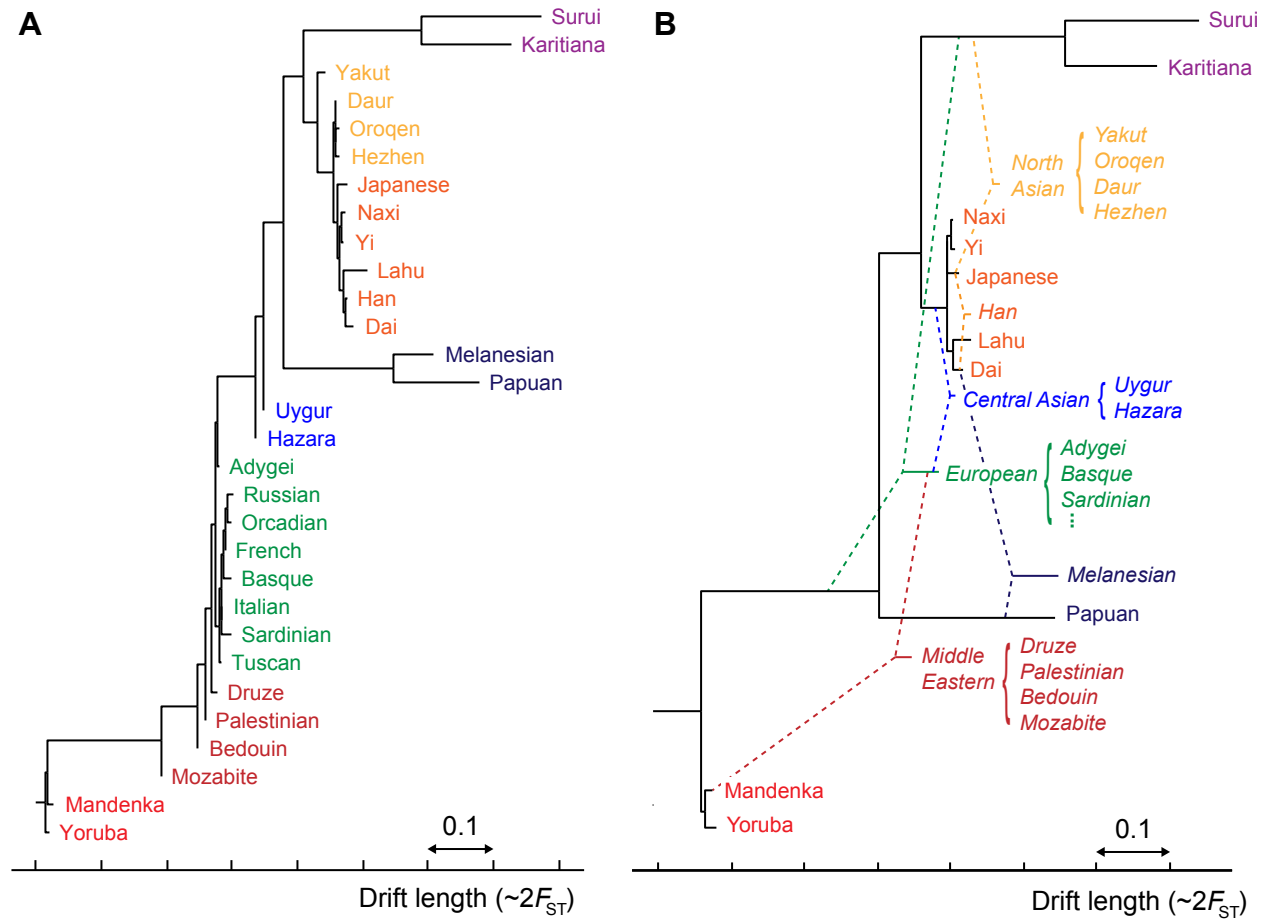


Figure 3. Aggregate phylogenetic trees of HGDP populations with and without admixture. (A) A simple neighbor-joining tree on the 30 populations for which *MixMapper* produced high-confidence results. This tree is analogous to the one given by Li et al. (2008, Figure 1B), and the topology is very similar (for the sake of comparison, we use the same color scheme). (B) Results from *MixMapper*. The populations appear in roughly the same order, but the majority are inferred to be admixed, as represented by dashed lines (cf. Pickrell and Pritchard (2012) and Figure S2). The admixed populations can be grouped into six categories, as indicated. Note that drift units are not additive, so branch lengths should be measured individually.

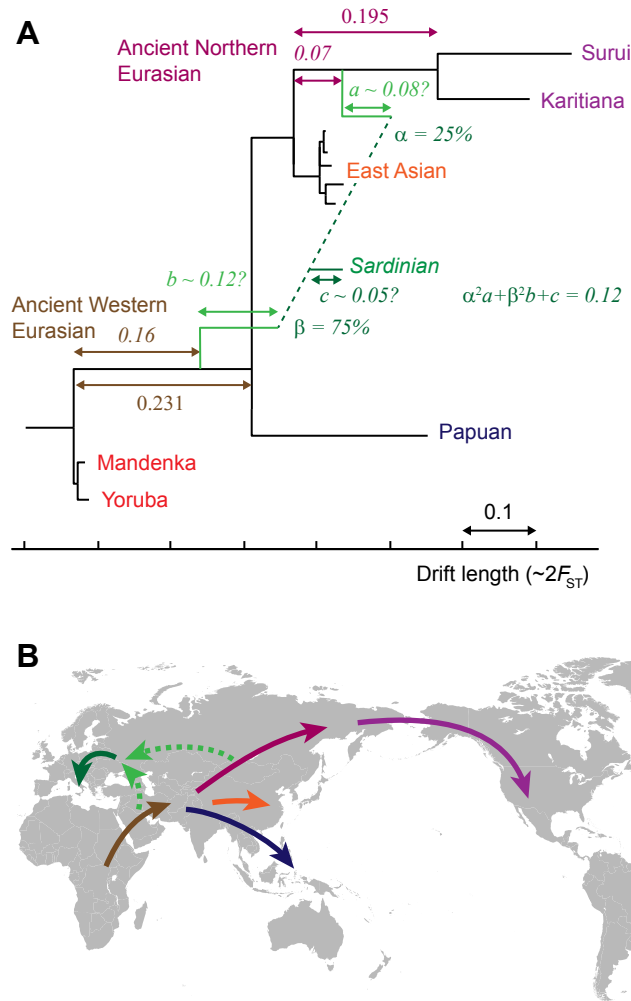


Figure 4. Inferred ancient admixture in Europe. (A) Detail of the inferred ancestral admixture for Sardinians (other European populations are similar). One mixing population splits from the unadmixed tree along the common ancestor branch of Americans (“Ancient Northern Eurasian”) and the other along the common ancestor branch of all non-Africans (“Ancient Western Eurasian”). Median parameter values are shown; 95% bootstrap confidence intervals can be found in Table 1. The branch lengths a , b , and c are confounded, so we show a plausible combination. (B) Map showing a sketch of possible directions of movement of ancestral populations. Colored arrows correspond to labeled branches in (A).

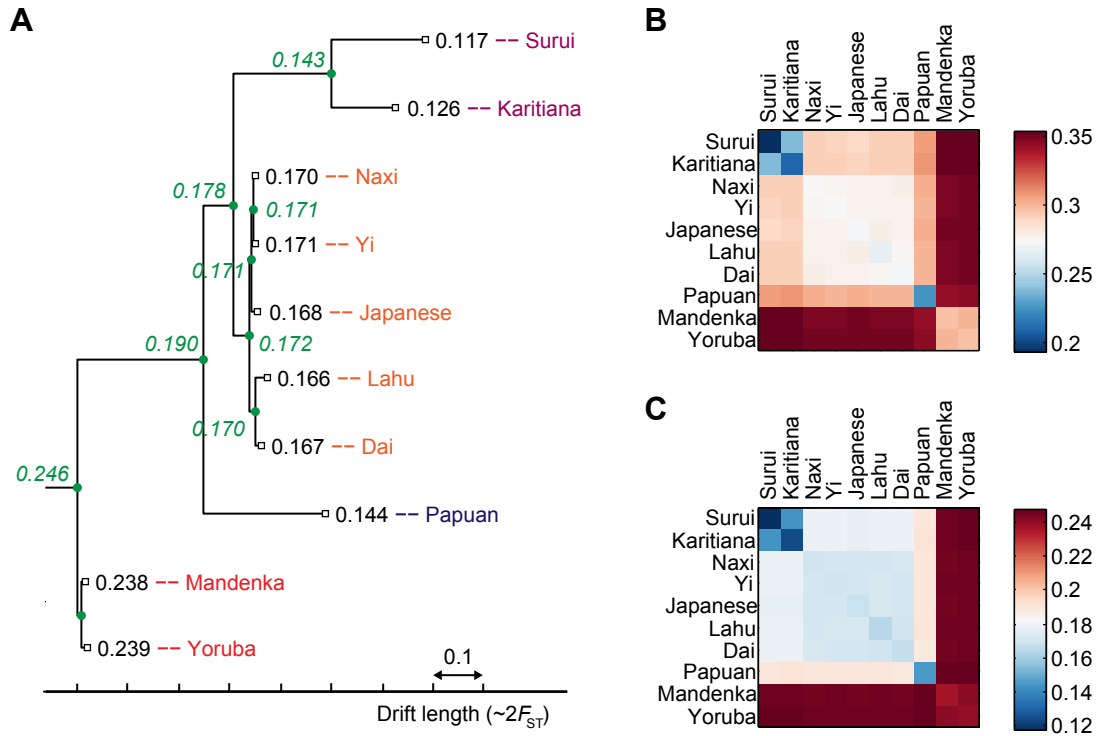


Figure 5. Ancestral heterozygosity imputed from original Illumina vs. San-ascertained SNPs. (A) The 10-population unadmixed tree with estimated average heterozygosities using SNPs from Panel 4 (San ascertainment) of the Affymetrix Human Origins array (Patterson et al., 2012). Numbers in black are direct calculations for modern populations, while numbers in green are inferred values at ancestral nodes. (B, C) Computed ancestral heterozygosity at the common ancestor of each pair of modern populations. With unbiased data, values should be equal for pairs having the same common ancestor. (B) Values from a filtered subset of about 250,000 SNPs from the published Illumina array data (Li et al., 2008). (C) Values from the Human Origins array excluding SNPs in gene regions.

Tables

Table 1. Mixture parameters for Europeans.

AdmixedPop	# rep	α	Branch1Loc (Anc. N. Eurasian)	Branch2Loc (Anc. W. Eurasian)	MixedDrift
Adygei	500	0.254-0.461	0.033-0.078 / 0.195	0.140-0.174 / 0.231	0.077-0.092
Basque	464	0.160-0.385	0.053-0.143 / 0.196	0.149-0.180 / 0.231	0.105-0.121
French	491	0.184-0.386	0.054-0.130 / 0.195	0.149-0.177 / 0.231	0.089-0.104
Italian	497	0.210-0.415	0.043-0.108 / 0.195	0.137-0.173 / 0.231	0.092-0.109
Orcadian	442	0.156-0.350	0.068-0.164 / 0.195	0.161-0.185 / 0.231	0.096-0.113
Russian	500	0.278-0.486	0.045-0.091 / 0.195	0.146-0.181 / 0.231	0.079-0.095
Sardinian	480	0.150-0.350	0.045-0.121 / 0.195	0.146-0.176 / 0.231	0.107-0.123
Tuscan	489	0.179-0.431	0.039-0.118 / 0.195	0.137-0.177 / 0.231	0.088-0.110

Mixture parameters from *MixMapper* for modern-day European populations (cf. Patterson et al. (2012)). All eight are nearly unanimously optimized as a mixture between populations related to the “ancient northern Eurasian” and “ancient western Eurasian” branches in the unadmixed tree (see Figure 4A). Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches; α is the proportion of ancestry from the “ancient northern Eurasian” side; MixedDrift is the sum of drift lengths $\alpha^2 a + (1 - \alpha)^2 b + c$; and # rep is the number of bootstrap replicates (out of 500) placing the mixture between these two branches. All ranges shown are 95% bootstrap confidence intervals. See Figure 2A for an illustration of the parameters.

Table 2. Mixture parameters for other populations modeled as two-way admixtures.

AdmixedPop	Branch1 + Branch2	# rep	α	Branch1Loc	Branch2Loc	MixedDrift
Daur	Anc. N. Eurasian + Japanese	350	0.067-0.276	0.008-0.126 / 0.195	0.006-0.013 / 0.016	0.006-0.015
	Suruí + Japanese	112	0.021-0.058	0.008-0.177 / 0.177	0.005-0.010 / 0.015	0.005-0.016
Hezhen	Anc. N. Eurasian + Japanese	411	0.068-0.273	0.006-0.113 / 0.195	0.006-0.013 / 0.016	0.005-0.029
Oroqen	Anc. N. Eurasian + Japanese	410	0.093-0.333	0.017-0.133 / 0.195	0.005-0.013 / 0.015	0.011-0.030
	Karitiana + Japanese	53	0.025-0.086	0.014-0.136 / 0.136	0.004-0.008 / 0.016	0.008-0.026
Yakut	Anc. N. Eurasian + Japanese	481	0.494-0.769	0.005-0.026 / 0.195	0.012-0.016 / 0.016	0.030-0.041
Melanesian	Dai + Papuan	424	0.160-0.260	0.008-0.014 / 0.014	0.165-0.201 / 0.247	0.089-0.114
	Lahu + Papuan	54	0.155-0.255	0.003-0.032 / 0.032	0.167-0.208 / 0.249	0.081-0.114
Han	Dai + Japanese	440	0.349-0.690	0.004-0.014 / 0.014	0.008-0.016 / 0.016	0.002-0.006

Mixture parameters from *MixMapper* for non-European populations fit as two-way admixtures. See Figure 2A and the caption of Table 1 for descriptions of the parameters. Branch1 and Branch2 are the optimal split points for the mixing populations; branch choices are shown that occur for at least 50 of 500 bootstrap replicates.

Table 3. Mixture parameters for populations modeled as three-way admixtures.

AdmixedPop2	Branch3	# rep	α_2	Branch3Loc	MixedDrift1A	FinalDrift1B	MixedDrift2
Druze	Mandenka	330	0.963-0.988	0.000-0.009 / 0.009	0.081-0.099	0.022-0.030	0.004-0.013
	Yoruba	82	0.965-0.991	0.000-0.010 / 0.010	0.080-0.099	0.022-0.029	0.005-0.013
	Anc. W. Eurasian	79	0.881-0.966	0.041-0.158 / 0.232	0.092-0.118	0.000-0.024	0.010-0.031
Palestinian	Anc. W. Eurasian	294	0.818-0.901	0.031-0.104 / 0.231	0.093-0.123	0.000-0.021	0.007-0.022
	Mandenka	146	0.909-0.937	0.000-0.009 / 0.009	0.083-0.097	0.022-0.029	0.001-0.007
	Yoruba	53	0.911-0.938	0.000-0.010 / 0.010	0.077-0.098	0.021-0.029	0.001-0.008
Bedouin	Anc. W. Eurasian	271	0.767-0.873	0.019-0.086 / 0.231	0.094-0.122	0.000-0.022	0.012-0.031
	Mandenka	176	0.856-0.923	0.000-0.008 / 0.008	0.080-0.099	0.023-0.030	0.006-0.018
Mozabite	Mandenka	254	0.686-0.775	0.000-0.009 / 0.009	0.088-0.109	0.012-0.022	0.017-0.032
	Anc. W. Eurasian	142	0.608-0.722	0.002-0.026 / 0.232	0.103-0.122	0.000-0.011	0.018-0.035
	Yoruba	73	0.669-0.767	0.000-0.008 / 0.010	0.086-0.108	0.012-0.023	0.017-0.031
Hazara	Anc. East Asian	497	0.364-0.471	0.010-0.024 / 0.034	0.080-0.115	0.004-0.034	0.004-0.013
Uygur	Anc. East Asian	500	0.318-0.438	0.007-0.023 / 0.034	0.088-0.123	0.000-0.027	0.000-0.009

Mixture parameters from *MixMapper* for populations fit as three-way admixtures. In all cases one parent population splits from the (admixed) Sardinian branch and the other from Branch3. See Figure 2B and the caption of Table 1 for further descriptions of the parameters. Branch choices are shown that occur for at least 50 of 500 bootstrap replicates. The “Anc. East Asian” branch is the common ancestral branch of the five East Asian populations in the unadmixed tree (Dai, Japanese, Lahu, Naxi, and Yi).

Supporting Information

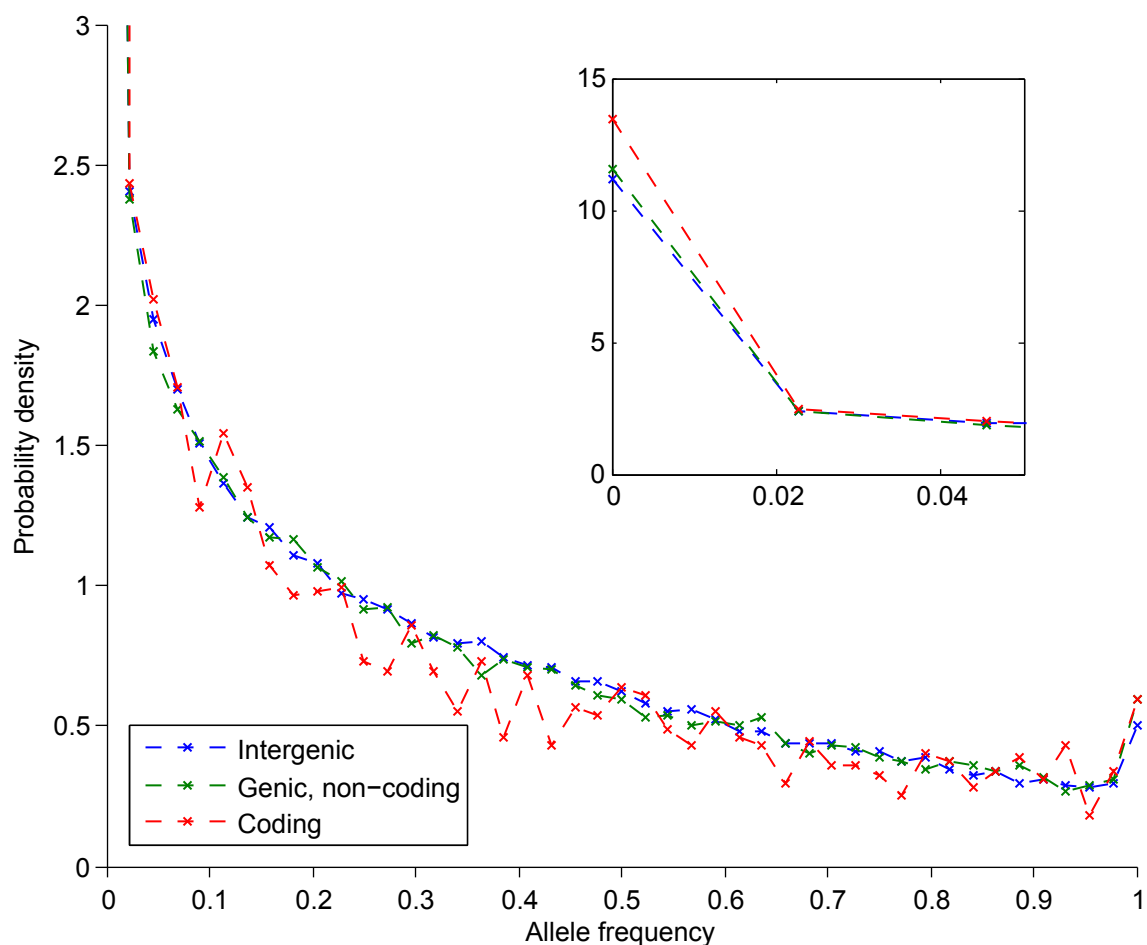


Figure S1. Comparison of allele frequency spectra within and outside gene regions. We divided the Panel 4 (San-ascertained) SNPs into three groups: those outside gene regions (101,944), those within gene regions but not in exons (58,110), and those within coding regions (3259). Allele frequency spectra restricted to each group are shown for the Yoruba population. Reduced heterozygosity within exon regions is evident, which suggests the action of purifying selection. (Inset) We observe the same effect in the genic, non-coding spectrum; it is less noticeable but can be seen at the edge of the spectrum.

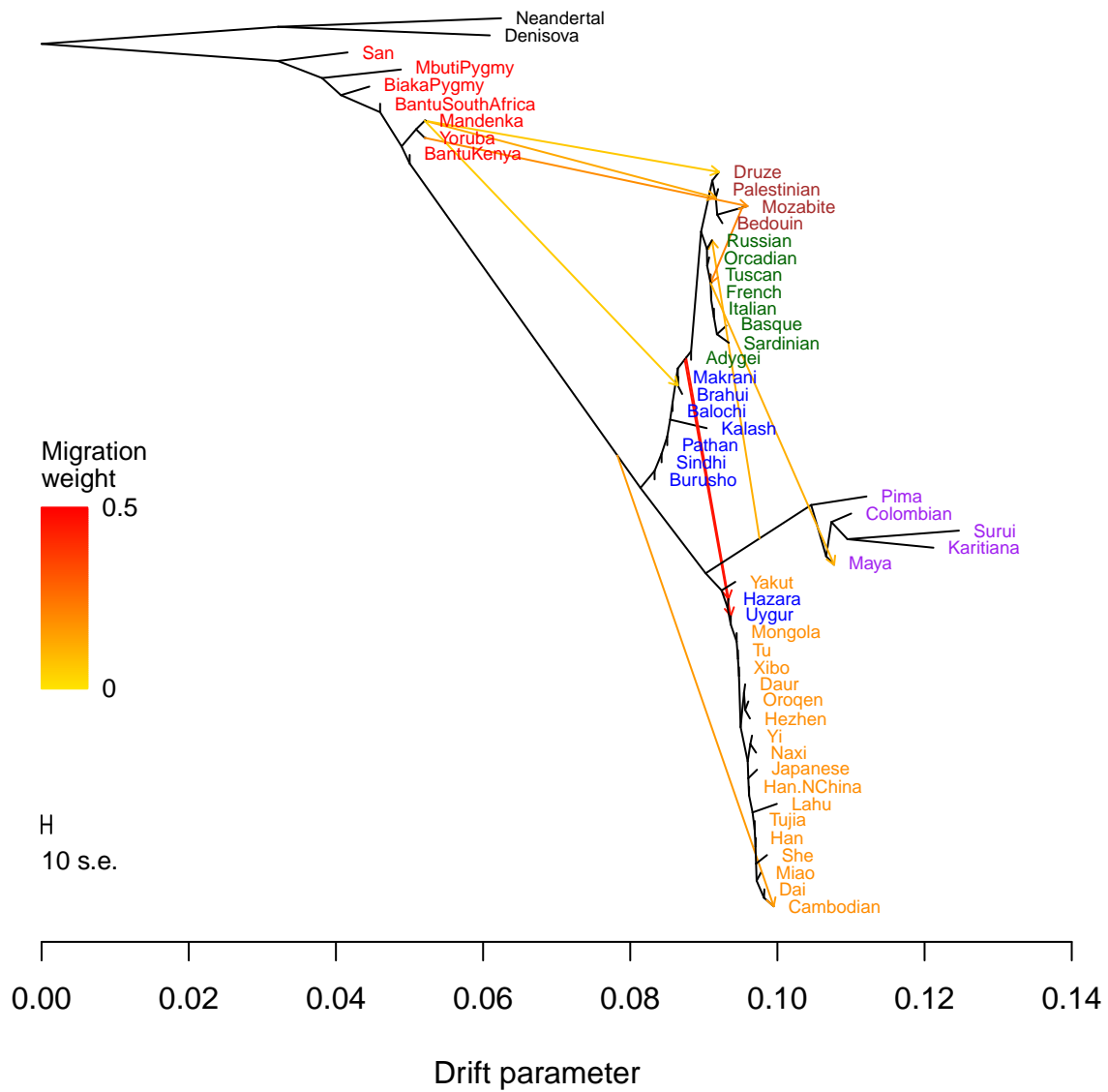
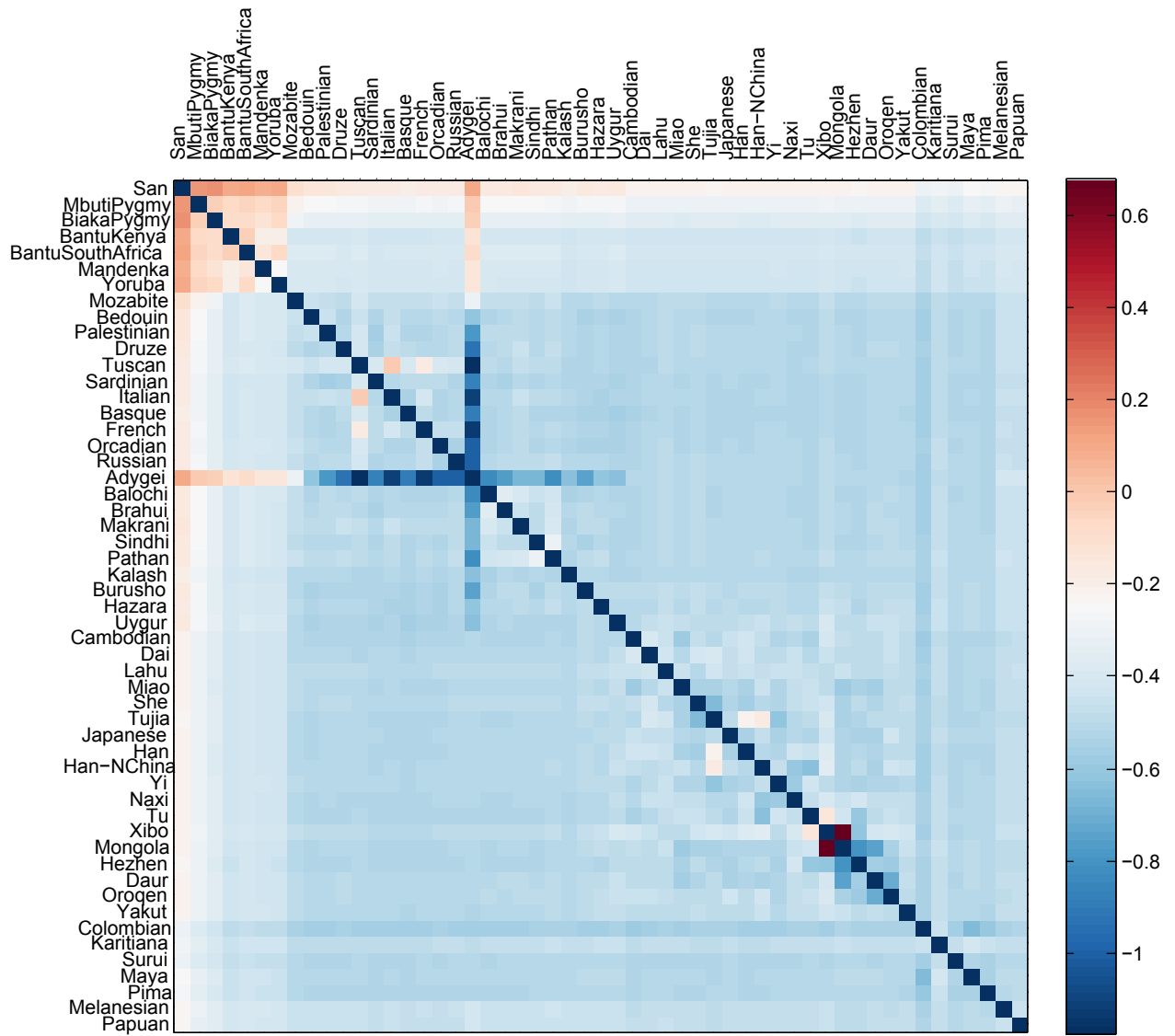


Figure S2. *TreeMix* results on the HGDP. Admixture graph for HGDP populations obtained with the *TreeMix* software, as reported in Pickrell and Pritchard (2012). Figure is reproduced from Pickrell and Pritchard (2012) with permission of the authors and under the Creative Commons Attribution License.



Log fold change in f_2 values (new array / original HGDP)

Figure S3. Comparison of f_2 distances computed using original Illumina vs. San-ascertained SNPs. The heat map shows the log fold change in f_2 values obtained from the original HGDP data (Li et al., 2008) versus the San-ascertained data (Patterson et al., 2012) used in this study.

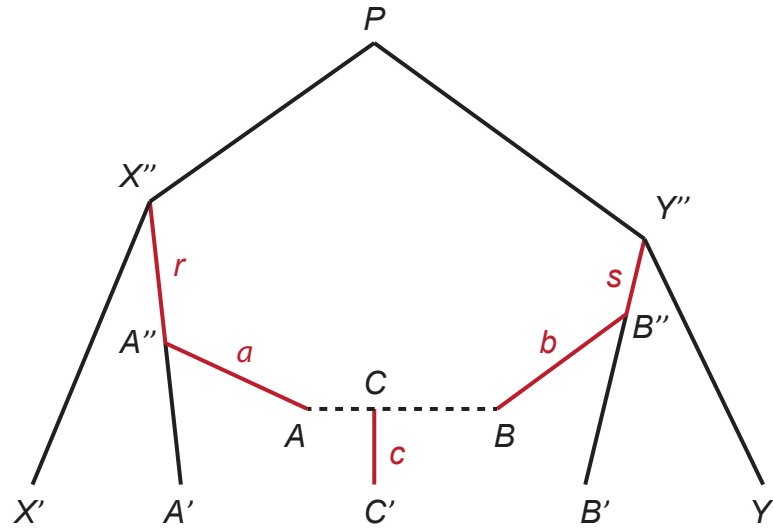


Figure S4. Schematic of part of an admixture tree. Population C is derived from an admixture of populations A and B with proportion α coming from A . The f_2 distances from C' to the present-day populations A' , B' , X' , Y' give four relations from which we are able to infer four parameters: the mixture fraction α , the locations of the split points A'' and B'' (i.e. r and s), and the combined drift $\alpha^2 a + (1 - \alpha)^2 b + c$.

Table S1. Mixture proportions for Sardinian and Basque from f_4 ratio estimation.

Test pop.	Asian pop.	American pop.	α
Sardinian	Dai	Karitiana	23.3 ± 6.3
Sardinian	Dai	Suruí	24.5 ± 6.7
Sardinian	Lahu	Karitiana	23.1 ± 7.0
Sardinian	Lahu	Suruí	24.7 ± 7.6
Basque	Dai	Karitiana	22.8 ± 7.0
Basque	Dai	Suruí	24.0 ± 7.6
Basque	Lahu	Karitiana	23.1 ± 7.4
Basque	Lahu	Suruí	24.7 ± 8.0

To validate the mixture proportions estimated by *MixMapper* for Sardinian and Basque, we applied f_4 ratio estimation. The fraction α of “ancient northern Eurasian” ancestry was estimated as $\alpha = f_4(\text{Papuan, Asian; Yoruba, European}) / f_4(\text{Papuan, Asian; Yoruba, American})$, where the European population is Sardinian or Basque, Asian is Dai or Lahu, and American is Karitiana or Suruí. Standard errors are from 500 bootstrap replicates. Note that this calculation assumes the topology of the ancestral mixing populations as inferred by *MixMapper* (Figure 4A).

Text S1. f -statistics and population admixture.

Here we include derivations of the allele frequency divergence equations solved by *MixMapper* to determine the optimal placement of admixed populations. These results were first presented in Reich et al. (2009) and Patterson et al. (2012), and we reproduce them here for completeness, with slightly different emphasis and notation. We also describe in the final paragraph (and in more detail in Methods) how the structure of the equations leads to a particular form of the system for a full admixture tree.

Our basic quantity of interest is the f -statistic f_2 , as defined in Reich et al. (2009), which is the squared allele frequency difference between two populations at a biallelic SNP. That is, at SNP locus i , we define

$$f_2^i(A, B) := (p_A - p_B)^2,$$

where p_A is the frequency of one allele in population A and p_B is the frequency of the allele in population B . This is the same as Nei's minimum genetic distance D_{AB} for the case of a biallelic locus (Nei, 1987). As in Reich et al. (2009), we define the unbiased estimator $\hat{f}_2^i(A, B)$, which is a function of finite population samples:

$$\hat{f}_2^i(A, B) := (\hat{p}_A - \hat{p}_B)^2 - \frac{\hat{p}_A(1 - \hat{p}_A)}{n_A - 1} - \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B - 1},$$

where, for each of A and B , \hat{p} is the empirical allele frequency and n is the total number of sampled alleles.

We can also think of $f_2^i(A, B)$ itself as the outcome of a random process of genetic history. In this context, we define

$$F_2^i(A, B) := E((p_A - p_B)^2),$$

the expectation of $(p_A - p_B)^2$ as a function of population parameters. So, for example, if B is descended from A via one generation of Wright-Fisher genetic drift in a population of size N , then $F_2^i(A, B) = p_A(1 - p_A)/2N$.

While $\hat{f}_2^i(A, B)$ is unbiased, its variance may be large, so in practice, we use the statistic

$$\hat{f}_2(A, B) := \frac{1}{m} \sum_{i=1}^m \hat{f}_2^i(A, B),$$

i.e., the average of $\hat{f}_2^i(A, B)$ over a set of m SNPs. As we discuss in more detail in Text S2, $\hat{f}_2^i(A, B)$ is not the same for different loci, meaning $\hat{f}_2(A, B)$ will depend on the choice of SNPs. However, we do know that $\hat{f}_2(A, B)$ is an unbiased estimator of the true average $f_2(A, B)$ of $f_2^i(A, B)$ over the set of SNPs.

The utility of the f_2 statistic is due largely to the relative ease of deriving equations for its expectation between populations on an admixture tree. The following derivations are borrowed from (Reich et al., 2009). As above, let the frequency of a SNP i in population X be p_X . Then, for example,

$$\begin{aligned} E(f_2^i(A, B)) &= E((p_A - p_B)^2) \\ &= E((p_A - p_P + p_P - p_B)^2) \\ &= E((p_A - p_P)^2) + E((p_P - p_B)^2) + 2E((p_A - p_P)(p_P - p_B)) \\ &= E(f_2^i(A, P)) + E(f_2^i(B, P)), \end{aligned}$$

since the genetic drifts $p_A - p_P$ and $p_P - p_B$ are uncorrelated and have expectation 0. We can decompose these terms further; if Q is a population along the branch between A and P , then:

$$\begin{aligned} E(f_2^i(A, P)) &= E((p_A - p_P)^2) \\ &= E((p_A - p_Q + p_Q - p_P)^2) \\ &= E((p_A - p_Q)^2) + E((p_Q - p_P)^2) + 2E((p_A - p_Q)(p_Q - p_P)) \\ &= E(f_2^i(A, Q)) + E(f_2^i(Q, P)). \end{aligned}$$

Here, again, $E(p_A - p_Q) = E(p_Q - p_P) = 0$, but $p_A - p_Q$ and $p_Q - p_P$ are not independent; for example, if $p_Q - p_P = -p_P$, i.e. $p_Q = 0$, then necessarily $p_A - p_Q = 0$. However, $p_A - p_Q$ and $p_Q - p_P$ are independent conditional on a single value of p_Q , meaning the conditional expectation of $(p_A - p_Q)(p_Q - p_P)$ is 0. By the double expectation theorem,

$$E((p_A - p_Q)(p_Q - p_P)) = E(E((p_A - p_Q)(p_Q - p_P)|p_Q)) = E(E(0)) = 0.$$

From $E(f_2^i(A, P)) = E(f_2^i(A, Q)) + E(f_2^i(Q, P))$, we can take the average over a set of SNPs to yield, in the notation from above,

$$F_2(A, P) = F_2(A, Q) + F_2(Q, P).$$

We have thus shown that f_2 distances are additive along an unadmixed-drift tree. This property is fundamental for our theoretical results and is also essential for finding admixtures, since, as we will see, additivity does not hold for admixed populations.

Given a set of populations with allele frequencies at a set of SNPs, we can use the estimator \hat{f}_2 to compute f_2 distances between each pair. These distances should be additive if the populations are related as a true tree. Thus, it is natural to build a phylogeny using neighbor-joining (Saitou and Nei, 1987), yielding a fully parameterized tree with all branch lengths inferred. However, in practice, the tree will not exactly be additive, and we may wish to try fitting some population C' as an admixture. To do so, we would have to specify six parameters (in the notation of Figure S4): the locations on the tree of A'' and B'' ; the branch lengths $f_2(A'', A)$, $f_2(B'', B)$, and $f_2(C, C')$; and the mixture fraction. These are the variables r , s , a , b , c , and α .

In order to fit C' onto an unadmixed tree (that is, solve for the six mixture parameters), we use the equations for the expectations $F_2(P, C')$ of the f_2 distances between C' and each other population P in the tree. Referring to Figure S4, with the point admixture model, the allele

frequency in C is $p_C = \alpha p_A + (1 - \alpha) p_B$. So, for a single locus, using additivity,

$$\begin{aligned}
E(f_2^i(A', C')) &= E((p_{A'} - p_{C'})^2) \\
&= E((p_{A'} - p_{A''} + p_{A''} - p_C + p_C - p_{C'})^2) \\
&= E((p_{A'} - p_{A''})^2) + E((p_{A''} - \alpha p_A - (1 - \alpha) p_B)^2) + E((p_C - p_{C'})^2) \\
&= E(f_2^i(A', A'')) + \alpha^2 E(f_2^i(A'', A)) \\
&\quad + (1 - \alpha)^2 E(f_2^i(A'', B)) + E(f_2^i(C, C')).
\end{aligned}$$

Averaging over SNPs, and replacing $E(f_2(A', C'))$ by the estimator $\hat{f}_2(A', C')$, this becomes

$$\begin{aligned}
\hat{f}_2(A', C') &= F_2(A', X'') - r + \alpha^2 a \\
&\quad + (1 - \alpha)^2 (r + F_2(X'', Y'') + s + b) + c \\
\implies \hat{f}_2(A', C') - F_2(A', X'') &= (\alpha^2 - 2\alpha)r + (1 - \alpha)^2 s + \alpha^2 a \\
&\quad + (1 - \alpha)^2 b + c + (1 - \alpha)^2 F_2(X'', Y'').
\end{aligned}$$

The quantities $F_2(X'', Y'')$ and $F_2(A', X'')$ are constants that can be read off of the neighbor-joining tree. Similarly, we have

$$\hat{f}_2(B', C') - F_2(B', Y'') = \alpha^2 r + (\alpha^2 - 1)s + \alpha^2 a + (1 - \alpha)^2 b + c + \alpha^2 F_2(X'', Y'').$$

For the outgroups X' and Y' , we have

$$\begin{aligned}
\hat{f}_2(X', C') &= \alpha^2 (c + a + r + F_2(X', X'')) \\
&\quad + (1 - \alpha)^2 (c + b + s + F_2(X'', Y'') + F_2(X', X'')) \\
&\quad + 2\alpha(1 - \alpha) (c + F_2(X', X'')) \\
&= \alpha^2 r + (1 - \alpha)^2 s + \alpha^2 a + (1 - \alpha)^2 b + c \\
&\quad + (1 - \alpha)^2 F_2(X'', Y'') + F_2(X', X'')
\end{aligned}$$

and

$$\hat{f}_2(Y', C') = \alpha^2 r + (1 - \alpha)^2 s + \alpha^2 a + (1 - \alpha)^2 b + c + \alpha^2 F_2(X'', Y'') + F_2(Y', Y'').$$

Assuming additivity within the neighbor-joining tree, any population descended from A'' will give the same equation (the first type), as will any population descended from B'' (the second type), and any outgroup (the third type, up to a constant and a coefficient of α). Thus, no matter how many populations there are in the unadmixed tree—and assuming there are at least two outgroups X' and Y' such that the points X'' and Y'' are distinct—the system of equations consisting of $E(f_2(P, C'))$ for all P will contain precisely enough information to solve for α , r , s , and the linear combination $\alpha^2 a + (1 - \alpha)^2 b + c$. We also note the useful fact that for a fixed value of α , the system is linear in the remaining variables.

Text S2. Heterozygosity and drift lengths.

One disadvantage to building trees with f_2 statistics is that the values are not in easily interpretable units. For a single locus, the f_2 statistic measures the squared allele frequency change between two populations. However, in practice, one needs to compute an average f_2 value over many loci. Since the amount of drift per generation is proportional to $p(1 - p)$, the expected frequency change in a given time interval will be different for loci with different initial frequencies. This means that the estimator \hat{f}_2 depends on the distribution of frequencies of the SNPs used to calculate it. For example, within an f_2 -based phylogeny, the lengths of non-adjacent edges are not directly comparable.

In order to make use of the properties of f_2 statistics for admixture tree building and still be able to present our final trees in more directly meaningful units, we will show now how f_2 distances can be converted into absolute drift lengths. Again, we consider a biallelic, neutral SNP in two populations, with no further mutations, under a Wright-Fisher model of genetic drift.

Suppose populations A and B are descended independently from a population P , and we have an allele with frequency p in P , $p_A = p + a$ in A , and $p_B = p + b$ in B . The (true) heterozygosities at this locus are $h_P^i = 2p(1 - p)$, $h_A^i = 2p_A(1 - p_A)$, and $h_B^i = 2p_B(1 - p_B)$. As above, we write \hat{h}_A^i for the unbiased single-locus estimator

$$\hat{h}_A^i := \frac{2n_A}{(n_A - 1)\hat{p}_A(1 - \hat{p}_A)},$$

\hat{h}_A for the multi-locus average of \hat{h}_A^i , and H_A^i for the expectation of h_A^i under the Wright-Fisher model (and similarly for B and P).

Say A has experienced t_A generations of drift with effective population size N_A since the split from P , and B has experienced t_B generations of drift with effective population size N_B . Then it is well known that $H_A^i = h_P^i(1 - D_A)$, where $D_A = 1 - (1 - 1/(2N_A))^{t_A}$, and

$H_B^i = h_P^i(1 - D_B)$. We also have

$$\begin{aligned}
H_A^i &= E(2(p + a)(1 - p - a)) \\
&= E(h_P^i - 2ap + 2a - 2ap - 2a^2) \\
&= h_P^i - 2E(a^2) \\
&= h_P^i - 2F_2^i(A, P),
\end{aligned}$$

so $2F_2^i(A, P) = h_P^i D_A$. Likewise, $2F_2^i(B, P) = h_P^i D_B$ and $2F_2^i(A, B) = h_P^i(D_A + D_B)$.

Finally,

$$H_A^i + H_B^i + 2F_2^i(A, B) = h_P^i(1 - D_A) + h_P^i(1 - D_B) + h_P^i(D_A + D_B) = 2h_P^i.$$

This equation is essentially equivalent to one in Nei (1987), although Nei interprets his version as a way to calculate the expected present-day heterozygosity rather than estimate the ancestral heterozygosity. To our knowledge, the equation has not been applied in the past for this second purpose.

In terms of allele frequencies, the form of h_P^i turns out to be very simple:

$$h_P^i = p_A + p_B - 2p_A p_B = p_A(1 - p_B) + p_B(1 - p_A),$$

which is the probability that two alleles, one sampled from A and one from B , are different by state. We can see, therefore, that this probability remains constant in expectation after any amount of drift in A and B . This fact is easily proved directly:

$$E(p_A + p_B - 2p_A p_B) = 2p - 2p^2 = h_P^i,$$

where we use the independence of drift in A and B .

Let $\hat{h}_P^i := (\hat{h}_A^i + \hat{h}_B^i + 2\hat{f}_2^i(A, B))/2$, and let h_P denote the true average heterozygosity in P over an entire set of SNPs. Since \hat{h}_P^i is an unbiased estimator of $(h_A^i + h_B^i + 2f_2^i(A, B))/2$, its expectation under the Wright-Fisher model is h_P^i . So, the average \hat{h}_P of \hat{h}_P^i over a set of SNPs is an unbiased (and potentially low-variance) estimator of h_P . If we have already constructed a phylogenetic tree using pairwise f_2 statistics, we can use the inferred branch length $\hat{f}_2(A', P)$ from a present-day population A to an ancestor P in order to estimate \hat{h}_P more directly as $\hat{h}_P = \hat{h}_A + 2\hat{f}_2(A, P)$. This allows us, for example, to estimate heterozygosities at intermediate points along branches or in the ancestors of present-day admixed populations.

The statistic \hat{h}_P is interesting in its own right, as it gives an unbiased estimate of the heterozygosity in the common ancestor of any pair of populations (for a certain subset of the genome). For our purposes, though, it is most useful because we can form the quotient

$$\hat{d}_A := \frac{2\hat{f}_2(A, P)}{\hat{h}_P},$$

where the f_2 statistic is inferred from a tree. This statistic \hat{d}_A is not exactly unbiased, but by the law of large numbers, if we use many SNPs, its expectation is very nearly

$$E(\hat{d}_A) \approx \frac{E(2\hat{f}_2(A, P))}{E(\hat{h}_P)} = \frac{h_P D_A}{h_P} = D_A,$$

where we use the fact that D_A is the same for all loci. Thus \hat{d} is a simple, direct, nearly unbiased moment estimator for the drift length between a population and one of its ancestors. This allows us to convert branch lengths from f_2 distances into absolute drift lengths, one branch at a time, by inferring ancestral heterozygosities and then dividing.

For a terminal admixed branch leading to a present-day population C' with heterozygosity $\hat{h}_{C'}$, we divide twice the inferred mixed drift $c_1 = \alpha^2 a + (1 - \alpha)^2 b + c$ (Figure 2) by the heterozygosity $\hat{h}_{C'}^* := \hat{h}_{C'} + 2c_1$. This is only an approximate conversion, since it utilizes a common value $\hat{h}_{C'}^*$ for what are really three disjoint branches, but the error should be very small with short drifts.

An alternative definition of \hat{d}_A would be $1 - \hat{h}_A/\hat{h}_P$, which also has expectation (roughly) D_A . In most cases, we prefer to use the definition in the previous paragraph, which allows us to leverage the greater robustness of the f_2 statistics, especially when taken from a multi-population tree.

We note that this estimate of drift lengths is similar in spirit to the widely-used statistic F_{ST} . For example, under proper conditions, the expectation of F_{ST} among populations that have diverged under unadmixed drift is also $1 - (1 - 1/(2N_e))^t$ (Nei, 1987). When F_{ST} is calculated for two populations at a biallelic locus using the formula $(\Pi_D - \Pi_S)/\Pi_D$, where Π_D is the probability two alleles from different populations are different by state and Π_S is the (average) probability two alleles from the same population are different by state (as in Reich et al. (2009) or the measure G'_{ST} in Nei (1987)), then this F_{ST} is exactly half of our \hat{d} . As a general rule, drift lengths \hat{d} are approximately twice as large as values of F_{ST} reported elsewhere.